The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies

de Wit, Jan; Schodde, Thorsten; Willemsen, Bram; Bergmann, Kirsten; de Haas, Mirjam; Kopp, Stefan; Krahmer, Emiel; Vogt, Paul

Published in:
Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies

Jan de Wit
TiCC*
Tilburg University
j.m.s.dewit@uvt.nl

Thorsten Schodde
Faculty of Technology, CITEC‖
Bielefeld University
tschodde@techfak.uni-bielefeld.de

Bram Willemsen
TiCC*
Tilburg University
b.willemsen@uvt.nl

Kirsten Bergmann
Faculty of Technology, CITEC‖
Bielefeld University
kirsten.bergmann@uni-bielefeld.de

Mirjam de Haas
TiCC*
Tilburg University
mirjam.dehaas@uvt.nl

Stefan Kopp
Faculty of Technology, CITEC‖
Bielefeld University
skopp@techfak.uni-bielefeld.de

Emiel Krahmer
TiCC*
Tilburg University
e.j.krahmer@uvt.nl

Paul Vogt
TiCC*
Tilburg University
p.a.vogt@uvt.nl

ABSTRACT

This paper presents a study in which children, four to six years old, were taught words in a second language by a robot tutor. The goal is to evaluate two ways for a robot to provide scaffolding for students: the use of iconic gestures, combined with adaptively choosing the next learning task based on the child's past performance. The results show a positive effect on long-term memorization of novel words, and an overall higher level of engagement during the learning activities when gestures are used. The adaptive tutoring strategy reduces the extent to which the level of engagement is diminishing during the later part of the interaction.

KEYWORDS

Language tutoring; Robotics; Education; Human-Robot Interaction; Bayesian Knowledge Tracing; Non-verbal communication

ACM Reference Format:
Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In *HRI '18: 2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3171221.3171277

1 INTRODUCTION

Robots show great potential in the field of education [24]. Embodied agents in the form of humanoid robots, in particular, may deliver educational content for various subjects in ways similar to human tutors. The main advantage of using such a robot compared to traditional learning tools is its physical presence in the referential world of the learner [20]. The human-like appearance and presence in the physical environment may facilitate interactions that are, to some extent, similar to the ways in which human teachers would communicate with their students. Care should be taken, however, to design for the correct amount of social behavior, so as to avoid distracting students from the task at hand [16].

When designing such interactions, we can draw upon ways in which human teachers give contingent support to students in their learning activities. For instance, particularly in one-on-one tutoring situations, teachers tend to adjust the pace and difficulty of learning tasks based on the past development and current skill set of the student [29]. For example, teachers may help by scaffolding, taking the initial knowledge base as a starting point and trying to optimize the learning gain by choosing the hardest task to perform that still lies within the zone of proximal development [32] of the student.

The use of gestures that coincide with speech is another way for teachers to provide scaffolding, particularly when the concepts which the gestures refer to are not yet mastered by the student [1]. For instance, when teaching a second language (L2), gestures can help to ground an unknown word in the target language by linking it iconically or indexically to a real world concept. Such a facilitating effect on word learning has been found for imitating gestures of a virtual avatar [2]. However, it is an open question if the embodied presence of a robot can be exploited to support language learning through a robot's gesturing, and if so, what kind of gestures would have a positive impact.

In this paper, we present the results of an experiment conducted to explore how these two tools for scaffolding the learning of language — choosing the task that yields the greatest potential learning gain for a particular student and the use of appropriate co-speech gestures — carry over to a humanoid robot. Both were combined

HRI '18, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171277

*Tilburg center for Cognition and Communication
‖Cluster of Excellence Cognitive Interaction Technology

in one study to better estimate what the relative importance of the respective techniques is, while keeping all other factors constant, and to find out whether the benefits of the two strategies can potentially reinforce or impede each other. The techniques were implemented and tested in a one-on-one tutoring system where children, four to six years old, play a game with a robot to learn an L2. In the next section, we briefly present the approaches taken to realize the adaptive tutoring along with co-speech gesturing of the robot. We then describe the experimental methodology, before reporting and discussing the results obtained.

2 BACKGROUND

2.1 Adaptive Bayesian Knowledge Tracing

A robot tutor that personalizes the learning experience for individual students has been shown to have a positive effect on performance [19]. This robot is also perceived as smarter or more intelligent and less distracting or annoying. In order to simulate the way human tutors tailor learning activities and difficulty levels to a particular student, an adaptive tutoring system would have to measure and track the knowledge level of the student. Often the knowledge is traced skill-wise, where in the case of language learning, the mastery of particular words or phrases in the target language is represented probabilistically (e.g., [11]). This approach yields promising results, but it lacks flexibility because of the need to define domain-specific distance metrics to choose the next skill. Others have used Dynamic Bayesian Networks to represent the learner's knowledge about a skill, conditioned on the past interaction and taking into account skill interdependencies [14]. This approach requires detailed knowledge about the learning domain to model those interdependencies and their parameters. Recently, Spaulding et al. [27] used a simpler approach based on Bayesian Knowledge Tracing (BKT) [6]. The general BKT model consists of latent variables S^t representing the extent to which the system believes a particular skill to be mastered by the student. The belief state of the system is updated based on observed variables O^t, which correspond to the result of a learning action (e.g., correctly or incorrectly answering a question), while accounting for possible cases of guessing $p(guess)$ and slipping $p(slip)$ during the answer process. It was shown that this model outperforms traditional approaches for tracing the knowledge state in learning interactions, and that it can be easily extended to, for example, incorporate the emotional state of a child. In previous work [26], we have extended the basic BKT with action nodes to also model the tutor's decision-making based on current beliefs about the student's knowledge state (see Figure 1). Additionally, we employed a latent variable S that can attain discrete values for each skill, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%). This allows for quantifying the robot's uncertainty about a learner's skills as well as the impact of tutoring actions on future observations and skills.

This so-called *Adaptive* Bayesian Knowledge Tracing (A-BKT) approach can be used to choose the next skill from which the learner will most likely benefit, by estimating the greatest expected knowledge gains. It tries to maximize the belief of each skill while also balancing over all skills and not teaching a particular skill over and over again, even if the answer to the task was wrong and the skill belief is the lowest. The system does not only allow to choose the best skill to address next, but also the action to be used for scaffolding the learning of this skill. In this context, actions can be, for example, different types of exercises, pedagogical acts, or task difficulties. For the sake of simplicity, three task difficulties have been established (easy, medium, hard) to address a skill and to find the best action for a given skill.

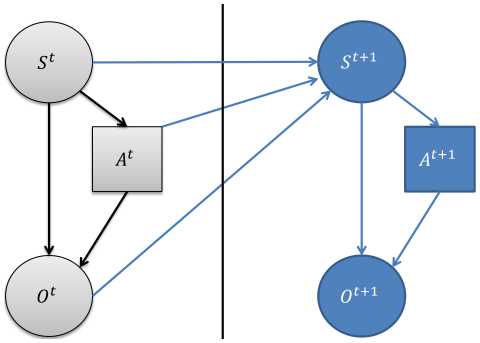

Figure 1: Dynamic Bayesian Network for BKT (taken from [26], with permission): with the current skill-belief the robot chooses the next skill S^t and action A^t for time step t and observes O^t as response from the user.

The goal of this strategy is to create a feeling of flow which can lead to better learning results [7]. It strives not to overburden the learner with tasks that would be too difficult nor to bore them with tasks that would be too easy, both of which may lead to disengagement and thus hamper the learning. Note that this approach is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [18]. The implementation of A-BKT used in the current study is identical to the one used previously in [26]. However, it has not yet been evaluated with children nor in conjunction with other techniques that might affect action difficulty (such as gestures). Furthermore, its impact on student engagement has not been explored previously.

2.2 Gestures

Iconic gestures elicit a mental image that corresponds directly, either in form or execution, to the concept or action that is being described verbally at the same time [23]. For example, a flying bird could be depicted by stretching both arms sideways and moving them up and down. Studies have shown that iconic gestures, when performed by a human teacher, may aid the acquisition of L2 vocabularies [8, 15, 21, 28]. Hald et al. [12] provide an overview of how gestures can contribute to learning an L2. They propose that gestures might have a 'grounding' effect by linking existing perceptual and motor experiences to a new word. This is expected to result in a richer mental representation. Research by Rowe et al. [25] shows that gender, language background, and level of experience in the native language (L1) influence the extent to which gestures can contribute to L2 learning. The positive effects of gestures hold true for younger students as well; in fact, gestures are suggested to be a crucial part of communication with children [13]. It has also been

shown that gestures help not only to acquire knowledge, but also to retain it over time [5].

Previous research has explored the use of gestures by virtual agents (e.g., [2]) and robots (e.g., [30]), finding similar, positive effects on memory performance when gestures are produced by an artificial embodied agent compared to a human tutor. While humans tend to spontaneously perform and time their gestures, they will often need to be manually designed and coordinated with speech for the robot. Due to its limited degrees of freedom, however, the robot is unable to perform motions with the same level of detail, finesse, and accuracy as a human. This may lead to a loss in meaning when human gestures are being translated directly to the robot, indicating a need for alternative gestures. As a concrete example, the SoftBank Robotics NAO robot that was used in this case is unable to move its three fingers individually, preventing it from performing pointing gestures or finger-counting. However, research suggests that iconic gestures are almost as comprehensible when performed by a robot, compared to a human [4].

3 METHODOLOGY

An experiment was conducted to investigate the effect of using iconic gestures and an adaptive tutoring strategy on children's acquisition of L2 vocabularies, with the intention of answering the following three hypotheses:

> H1: There is a greater learning gain when target words are accompanied by iconic gestures during training, than in the case of not using gestures.
> H2: There is a reduced knowledge decay when target words are accompanied by iconic gestures during training, than in the case of not using gestures.
> H3: There is a greater learning gain when target words are presented in an adaptive order during training, based on the knowledge state of the child, than when target words are randomly introduced.

These hypotheses rely upon the underlying assumption that children are able to acquire new L2 words during a single session with a robot tutor, regardless of experimental conditions; this assumption was also put to the test.

The experiment had a 2 (adaptive versus non-adaptive) x 2 (gestures versus no gestures) between-subjects design. In the two conditions with the adaptive tutoring strategy, the A-BKT system described in Section 2.1 was used to select the target word for each round, based on the believed knowledge state of the child. In practice, this meant that children would be presented with a particular target word more frequently if they had answered it incorrectly in the past, thereby changing the number of times each target word occurred during training, although each target word was guaranteed to occur at least once. Other conditions had a random selection, where each of the six target words would always be presented five times, in a randomized order, for a total of thirty rounds. In the gesture conditions, whenever a target word was introduced in the L2 it was accompanied by an iconic gesture (as shown in Figure 2). All conditions had the robot standing up and in "breathing" mode, which meant that it slowly shifted its weight from one leg to the other and had a slight movement in its arms to simulate breathing.

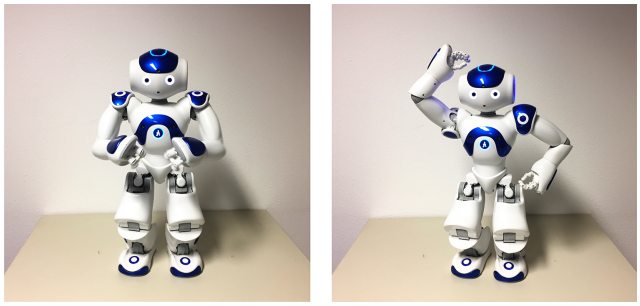

Figure 2: Examples of the stroke of two iconic gestures performed by the robot (taken from [9], with permission). Left: imitating a *chicken* by simulating the flapping of its wings; right: imitating a *monkey* by scratching head and armpit.

3.1 Participants

Participants were 61 children, with an average age of 5 years and 2 months ($SD = 7\,months$), 32 girls. They were recruited from primary schools in the Netherlands, by first contacting schools and then sending out an information letter together with a consent form through the schools to the parents of children that satisfied the age limit of four to six years. Only native Dutch children with Dutch as their L1 are included in the evaluation, although all 99 children that had signed up were allowed to participate in the experiment. The children were randomly assigned to conditions, while taking into account a balance in age and gender.

3.2 Materials

The aim of the tutoring interaction was to teach children six animal names in English: bird, chicken, hippo, horse, ladybug, and monkey. These specific words were chosen because the Dutch words are distinctly different from their English translations and because it was possible to create uniquely defining iconic gestures for them.

The SoftBank Robotics NAO robot was used, which was standing in front and slightly to the right of the child. After an experimenter had filled in the name of the child and pressed the start button, the experiment ran fully autonomously. Two experimenters were always present, where one would take care of getting the child from the classroom and explaining the procedure of the experiment, while the other would set up the system. To avoid having the child seek them out for feedback, the experimenters would announce that they would be occupied. The child was asked to sit on pillows, close to the tablet which was raised on a box and slightly tilted. Two cameras were used to record the interaction, one facing the front of the child and one at an angle from the side. The basic setup is shown in Figure 3, although it differed slightly between locations due to the layout of the rooms. In the condition with gestures every occurrence of the target word in L2, except when giving feedback, was accompanied by the matching iconic gesture (see Figure 2). The gesture was timed in such a way that the pronunciation of the target word would coincide with the stroke of the gesture, i.e., the accented phase that is most related to the meaning. A perception study was conducted to evaluate the quality of the gestures [9], where 14 participants were shown video recordings of all six gestures

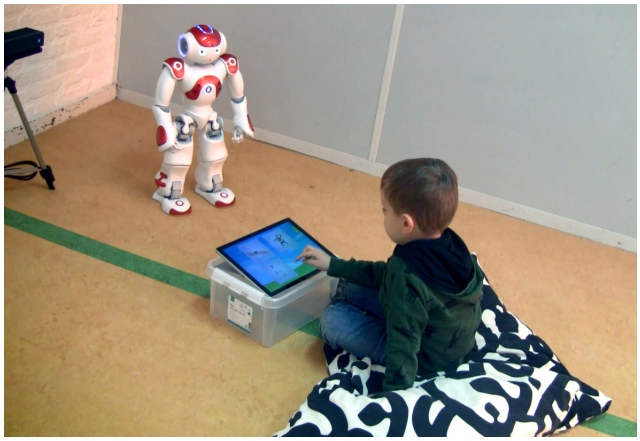

Figure 3: The setup for the experiments.

performed by the robot and then asked to indicate which out of the six target words corresponds to each particular recording. Based on the results of this study, each gesture was deemed to be sufficiently unique to distinguish between the six target words.

The adaptive tutoring system starts with medium (0.5) confidence for all target words, a value associated with two distractors during training. Each distractor is a false answer to a task, an image belonging to one of the five other target words. In the random conditions, since there is no knowledge tracing the difficulty was always set to medium (two distractors). The tablet was used to get input from the child, because speech recognition does not work reliably with children [17]. This is also why only comprehension and not production of the target words is evaluated. An example of what the tablet screen would look like is shown in Figure 5. The images used during training belong to a different set of images than the ones used for the pre-test and post-tests. The set of images used during training matches the gesture that the robot performs related to the animals, for example the image of the horse for the training stage (shown in Figure 5) also includes a rider because the robot shows the act of riding a horse as a gesture. The image that was used during the tests did not include a rider and the horse is standing still, facing the opposite direction (shown in Figure 4). In addition to changing the pose or context of the animals, colors also varied. Together with having a recorded voice in the tests instead of the robot's synthesized speech, this aims to verify whether children learn how the English words map to the concepts of the animals and their matching Dutch words, rather than to one specific image.

3.3 Procedure

Prior to partaking in the experiment, participants were introduced to the robot during a group introduction. This approach is inspired by the work of Vogt et al. [31] with the intention of lowering the anxiety of children in subsequent one-on-one interactions with the robot. The introduction consisted of a description of what the robot is like, including a background story and how it is similar to humans in some respects, and different in others. Together with the children (and sometimes teachers and experimenters) the robot performed dances, after which all children were presented with

Figure 4: The pre-test and post-tests on a laptop, using a recorded voice and a different set of images from those on the tablet.

Figure 5: The tablet during training, showing images corresponding to the target word and two distractors.

the opportunity to shake the robot's hand before putting it to bed. Introductory sessions were scheduled several days before the first participant was to take part in the experiment, allowing time for the children to process these new impressions.

Before starting the tutoring interaction, a pre-test was administered to gauge the level of prior knowledge with respect to the animal names in the L1 (Dutch) and L2 (English). This test was administered on a laptop, where images of all six animals were randomly positioned on the screen. A recording of a (bilingual) native speaker pronouncing one of the six animal names was played, after which the child was asked to click the corresponding image on the screen (Figure 4). This was done for all six target words, first in Dutch and then in English.

After completing the pre-tests, the child would go through each target word one by one, still using the laptop. This is done to give the children a first exposure to the correct mappings between target words and the concepts they refer to, to avoid turning the first rounds of learning with the robot into a guessing game. Because there is no feedback during the pre-tests, this also ensures that concepts are linked to the correct word, rather than having the child assume that their answers during the pre-tests were all correct. For each word, the image of the corresponding animal would be shown in the center of the screen and the laptop would play a recording by a (bilingual) native speaker saying: "Look, this is a [target in L2]. Do you see the [target in L2]? Click on the [target in L2]!"

The training stage of the experiment consisted of the child and robot playing thirty rounds of the game *I spy with my little eye*. The robot, acting as the spy, would pick one of six target words and call out: "I spy with my little eye...", followed by the chosen word in the L2. For this stage, children were assigned to one of four conditions:

(1) Random tutoring strategy, no gestures ($N = 16$)
(2) Random tutoring strategy, gestures ($N = 14$)
(3) Adaptive tutoring strategy, no gestures ($N = 15$)
(4) Adaptive tutoring strategy, gestures ($N = 16$)

Prior to playing the game, the robot explained the procedure and asked the child to indicate whether they understood by pressing either a green or a red smiley. If the red smiley is pressed, the interaction would pause and an experimenter would step in to provide any further explanations. After this introduction, there were two practice rounds: one in Dutch and one in English.

After the robot had "spied" an animal, a corresponding image was shown on the tablet along with a number of distractor images (Figure 5). The child was then asked to pick the image that matched the animal name that the robot had spied. The number of distractors was determined by the difficulty level of the round, which in the case of the adaptive conditions depended on the confidence that the system had in that the child knew this particular target word. A low confidence resulted in only one distractor, while a high confidence had three distractors.

Feedback to the task was given by both the tablet and the robot. The tablet highlighted the image selected by the participant, either with a green, happy smiley if the correct answer was provided or a red, sad smiley if the selected image was an incorrect answer. The robot then provided verbal feedback, which in the case of a correct answer consisted of a random pick out of six positive feedback phrases (e.g., "well done!"), followed by "The English word for [target in L1] is [target in L2]". In the case of negative feedback, the robot would say "That was a [chosen answer in L1], but I saw a [target in L2]. [Target in L2] is the English word for [target in L1]". Whenever an incorrect answer was given, the same round would be presented once more but at the easiest difficulty (with only one distractor: the image that was incorrectly chosen in the previous attempt). This, combined with additional exposures in the corrective feedback, means that the number of times each target word was presented in the L2 may vary between children, depending on how many rounds were answered incorrectly. After finishing thirty rounds of training with the robot, the child was asked to complete a post-test on the laptop. This test is identical to the pre-test that was administered at the start of the experiment, in L2. Finally, the post-test was repeated once more, at least one week after the experiment, to measure long-term retention of the newly acquired knowledge.

3.4 Analysis

Immediate learning gain was measured as the difference between the number of correct answers on the post-test, administered directly after the training stage, and the number of correct answers on the pre-test, taken prior to the tutoring interaction. Test scores were always between 0 and 6 because each target word was asked once in the L2. The post-test was administered once more, (at least) one week after the experiment. We then looked at the difference between this delayed test and the pre-test for long-term learning gain. Finally, we took the difference between the delayed test and the immediate post-test as a measure of knowledge decay. The design of these tests is described in more detail in Section 3.2.

Children's tasks during training were of varying task difficulty in the adaptive tutoring condition, with one to three distractor images. To account for these differences, as well as to allow a comparison with the post-test results (five distractor images), we mapped binary task success (1: correct response; 0: incorrect response) onto the span between 0.0 and 1.0 by subtracting a value of 0.2 for each of the potential five distractor images that was not provided, which would, for example, result in a score of 0.6 for a correct response in a task with three distractors. The total score during training was then divided by the number of rounds (30), resulting in a training performance value between 0.0 and 1.0 (Figure 7).

4 RESULTS

The average duration of the training stage of the experiment was 18:38 minutes (SD = 3:03). Including the introduction, pre-test, and post-test this amounted to a session length of roughly thirty minutes. To confirm whether children managed to learn any new words from a single tutoring interaction, regardless of strategy or the use of gestures, a paired-samples t-test was conducted to measure the difference between post-test and pre-test scores for all conditions combined. There was a significant difference between the scores on the pre-test ($M = 1.75$, $SD = 1.14$) and immediate post-test ($M = 2.85, SD = 1.61$), $t(60) = 5.23, p < .001$. The same analysis was conducted for the delayed post-test that was taken (at least) one week after the experiment. Results revealed a significant difference between the pre-test scores ($M = 1.75, SD = 1.14$) and the delayed post-test test scores ($M = 3.02, SD = 1.40$), $t(60) = 6.81, p < .001$. However, there was no significant difference between the delayed post-test and the immediate post-test, $t(60) = .92, p = .34$. This means that H2 is not supported by these results, since no significant decay was observed in any of the conditions.

To investigate the effects of the different conditions on training performance, a two-way ANOVA was carried out with tutoring strategy (adaptive versus non-adaptive) and the use of gestures (gestures versus no gestures) as independent variables and performance during training as the dependent variable (Figure 7). As described in Section 3.4, these scores are weighted by the number of distractors present and divided by 30 rounds, resulting in a value between 0.0 and 1.0. For the 30 rounds of training there was a main effect of gesture use, $F(1, 57) = 18.23, p < .001, \eta_p^2 = .24$, such that training with gestures led to higher score ($M = .38, SD = .09$) than learning without gestures ($M = .29, SD = .08$). Children in the adaptive condition achieved a higher score ($M = .36, SD = .12$) than children in the non-adaptive condition ($M = .32, SD = .06$), but the effect of tutoring strategy was not significant, $F(1, 57) = 3.62, p = .06, \eta_p^2 = .06$. There was a significant interaction effect between use of gestures and tutoring strategy, $F(1, 57) = 4.72, p = .03, \eta_p^2 = .08$. Without gesture use, there was no significant difference between tutoring strategies. When gestures were present, however, children in the adaptive condition turned out to perform better than those in the non-adaptive condition. Hence, children's learning outcome was best when gesture use and adaptive training were combined.

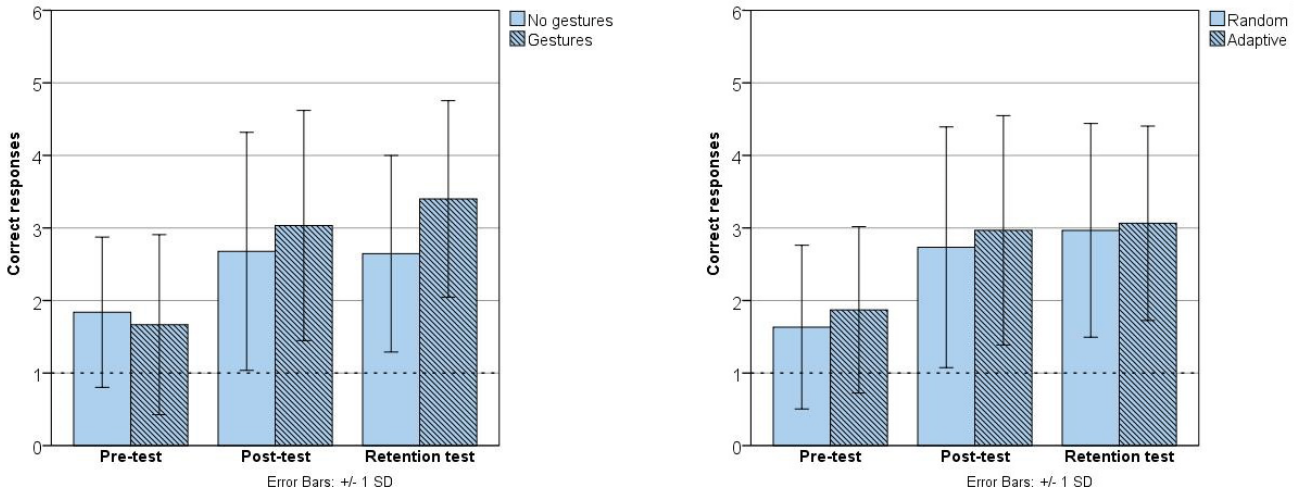

Figure 6: Test scores for the gesture vs no gesture conditions (left) and the adaptive vs random conditions (right).

Another two-way ANOVA was carried out to measure learning gain, with the difference score between the post-test results and the pre-test results as the dependent variable (Figure 6). There was no significant effect of tutoring strategy, $F(1, 57) < .001, p = .95, \eta_p^2 < .001$, or use of gestures, $F(1, 57) = 1.53, p = .22, \eta_p^2 = .03$. These results do not support H1 and H3 (greater learning gains when gestures and adaptive tutoring are used). The same two-way ANOVA with the difference score between results of the delayed post-test and the pre-test also did not give a significant effect of tutoring strategy, $F(1, 57) = .36, p = .55, \eta_p^2 = .006$, but there was a significant effect for use of gestures, $F(1, 57) = 6.11, p = .02, \eta_p^2 = .097$, indicating that the learning gain between pre-test and delayed post-test was greater when gestures were used during training ($M = 1.70, SD = 1.56$) than when no gestures were used ($M = .81, SD = 1.25$). Although this does not fully support H1 or H2, it does show a long-term learning gain when gestures are used during learning. No interaction effect was found, $F(1, 57) = .04, p = .84, \eta_p^2 \leq .001$.

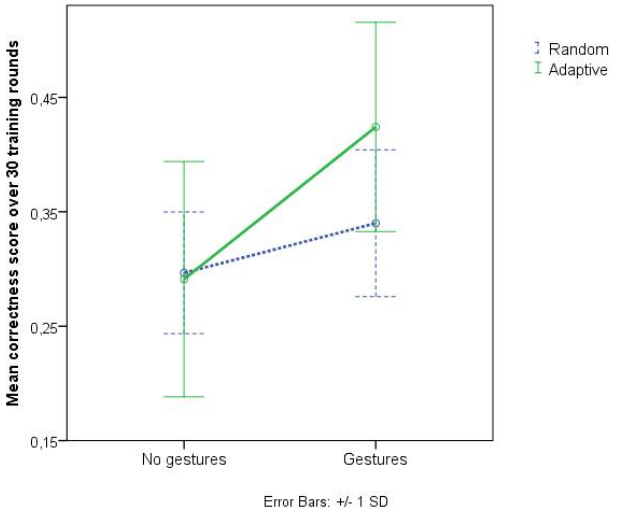

Figure 7: Interaction effects of gesture use and training strategy.

4.1 Evaluation of engagement

The engagement of the children during the training stage with the robot was examined to find out whether children became more disengaged with the tutoring tasks towards the end of the thirty rounds, and whether the application of an adaptive tutoring strategy and gestures would influence the change in engagement levels. This was done by asking 18 adult participants, without specific training in working with children, to rate video clips (without audio) of the children interacting with the robot. The choice for conducting a perception study with adults using video recordings of the experiment was made for two reasons: so that the training would not have to be interrupted for questions regarding the experience, thereby potentially influencing the engagement, and because it is difficult for children of a young age to reflect upon their experiences and verbalize these thoughts [22]. For each child, one clip was taken from the fifth round of training and one clip from the twenty-fifth round, to get observations that are close to the beginning and end of the training, but far enough from these actual moments to avoid short bursts of engagement when children realize the experiment is starting or finishing. The clips start right after the robot finishes introducing the task, i.e., the point at which the turn switches to the child to provide an answer. All clips then run for five seconds. One child that was excluded from the previous analysis because delayed post-test results were missing, was included for this part of the evaluation. However, data from one other child was missing, making the number of stimuli 122 (61 children, two clips each), with 14 to 16 children in each condition. Participants in the evaluation were asked to rate all 122 clips, randomly presented to them, on a scale from 1 (completely disengaged) to 7 (completely engaged). As a practice round, two clips of a child that was not included in the

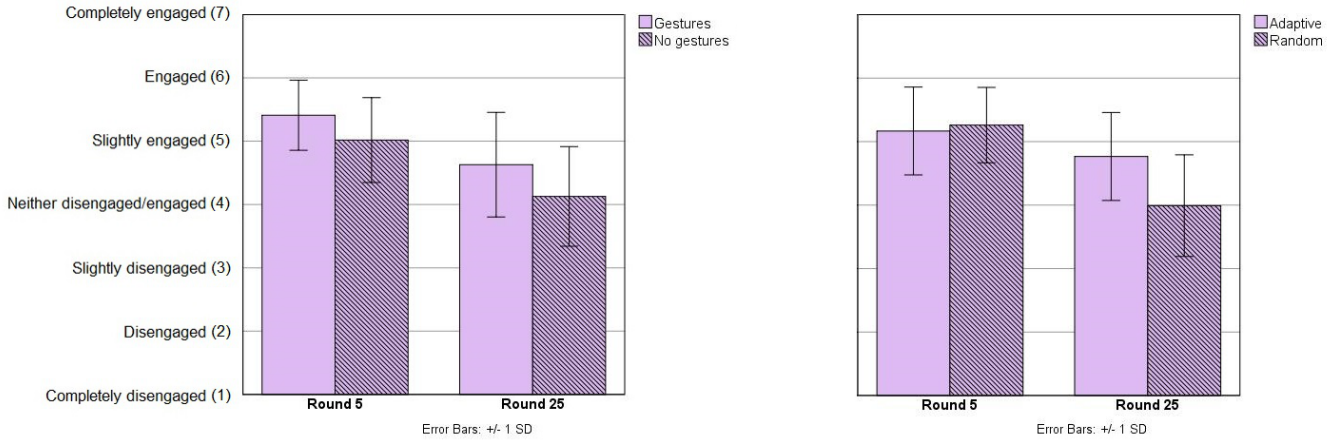

Figure 8: Rated engagement levels early and late in the training interaction for the gesture versus no gesture conditions (left) and the adaptive versus random conditions (right).

main experiment were presented, where one example was clearly engaged and the other was clearly not engaged. After this practice round, participants were told which features from the examples showed engagement (i.e., rapid response to the question, upright body posture, displaying joy after answering the question) and disengagement (i.e., slower response to the question, supporting the head by leaning on the arms, showing less interest in the task).

For each participant, the ratings were averaged over all children belonging to the same experimental condition, resulting in a total of eight average ratings (four conditions, each with fifth and twenty-fifth round). Figure 8 visualizes the data from the evaluation. Results from a paired-samples t-test showed that children were considered to be significantly less engaged in the twenty-fifth round ($M = 4.38, SD = .84$) than in the fifth round ($M = 5.21, SD = .64$), $t(71) = -12.09, p < .001$. Furthermore, a two-way ANOVA with tutoring strategy (adaptive versus non-adaptive) and gesture use (gestures versus no gestures) as factors showed no significant effect for the use of gestures, $F(1, 68) = 1.36, p = .25, \eta_p^2 = .02$, but there was a significant effect for tutoring strategy, $F(1, 68) = 86.26, p < .001, \eta_p^2 = .559$. The drop in engagement between round five and round twenty-five was less when an adaptive strategy was applied ($M = -.40, SD = .35$) than when words were randomly presented ($M = -1.27, SD = .44$). There was no interaction effect between gestures and tutoring strategies, $F(1, 68) = .01, p = .93, \eta_p^2 = .00$. The same analysis was conducted with the average engagement level of the fifth and twenty-fifth rounds combined, to get an idea of the overall engagement throughout the entire training session in different conditions. In this case the overall level of engagement was significantly higher in the gesture condition ($M = 5.02, SD = .63$) than in the condition without gestures ($M = 4.57, SD = .68$), $F(1, 68) = 8.75, p = .004, \eta_p^2 = .114$. There was also a significantly higher engagement when an adaptive strategy was used ($M = 4.97, SD = .67$) as opposed to a random tutoring strategy ($M = 4.63, SD = .67$), $F(1, 68) = 5.10, p = .03, \eta_p^2 = .07$. No interaction effect between the two factors was found, $F(1, 68) = .08, p = .78, \eta_p^2 = .001$.

5 DISCUSSION

The results presented above show that by spending a single tutoring interaction of about twenty minutes with a robot tutor, young children were able to acquire new words in an L2, regardless of the experimental condition, and were also able to retain this newly acquired knowledge for a prolonged period of time. Care was taken to design the pre-test and post-tests in such a way to be clearly distinct from the training session with the robot in terms of physical context (laptop versus tablet), voice, and characteristics of the images used, with the aim of getting a reliable measure of the attained knowledge. Results from the pre-test show that there is indeed a realistic amount of prior knowledge, on average above chance, presumably because some children have been exposed previously to the target words, for example in television programs. The observed number of correct answers on the immediate and delayed post-test are higher than on the pre-test, indicating the expected knowledge gain after engaging in learning activities. The scores on the post-test are lower than the number of correct answers towards the end of the training stage, which could show that indeed the test evaluates whether children acquire the underlying concepts, rather than simply being able to link a word being pronounced by the robot to one specific image (in some cases with the help of gestures that are not present in the tests). One potential point of improvement for the tests could be to introduce context when querying the target words, for example by using sentences rather than isolated words. Although explicitly instructed, children seemed not always aware that they were supposed to select the image corresponding to an *English* word, causing them to choose the animal with the most similar sounding name in Dutch instead (e.g., bird was often confused with the Dutch word 'paard').

When gestures were performed by the robot during training, there was a higher retention of newly acquired words after at least one week. This aligns with similar effects that were shown previously in the context of math with a human tutor [5] and indicates that these indeed carry over to a robot; a compelling finding that warrants future research into the intricacies of gesture use by humanoid robots. As mentioned by Hostetter [13] with respect to

human-human communication, it appears that gestures retain their positive effects on communication when they are scripted rather than being produced spontaneously. In this work, only iconic gestures are used that clearly relate to the concept they describe. Future work could investigate whether a similar contribution to learning gain is found when non-iconic gestures are used. Furthermore, the target words used in this experiment were chosen specifically such that matching gestures could be designed for the robot. It would be interesting to explore how well a broader range of gestures, describing various abstract and concrete concepts, could be performed by a robot as opposed to a human interlocutor. Finally, asking children to actually re-enact the gestures (e.g., as in [8, 28]), or to come up with their own gestures, might further increase the potential utility of gestures in learning due to the embodiment effect [10].

The test results regarding the adaptive tutoring system are currently inconclusive. This might be a result of the manner in which learning gain was measured, i.e., a quantification of newly acquired words — perhaps the adaptive system did not result in *more* words learned, but rather led to a more focused acquisition of exactly those words that the child found most difficult. The main remaining difference between the ways in which human teachers and the system presented here personalize content is that teachers tend to draw upon a memory that spans a longer period of time. In this experiment, the memory of the adaptive system was built up, and then applied, over the course of a single session. The system might come to fruition if there are multiple sessions with the same child, allowing the results of one session to become prior knowledge for the next one. It is also possible that the actions that the system performs based on the estimated knowledge levels of the child are too subtle. Currently, only the order and frequency of words is tailored, within the thirty rounds, and different levels of difficulty are represented by adding or removing one distractor image. Actions and difficulty levels could be more complex than that, for example by applying completely different tutoring strategies or games that might fit a particular child better. For the sake of this experiment, the number of rounds was fixed to thirty, but this session length might also be left up to the adaptive system to control. This would allow the interaction to end at the exact moment where the learning is 'optimal', i.e., a point at which the adaptive system thinks that the child has achieved his or her highest potential learning gain. A final avenue for improvement that is currently being pursued is to incorporate additional information about the affective state of the child. Some children might not be in the right mood to learn when they start, or their attention might fade during the interaction; rather than focusing only on the learning objectives the robot might want to engage in activities that work towards creating and maintaining the right atmosphere for learning.

We found it valuable to include the measure of children's engagement during the interaction. A higher level of engagement indicates increased motivation and willingness to learn [3]. Although students might succeed in simple word learning with limited engagement and the use of a low-level learning strategy, increased engagement could stimulate them to go beyond simple memorization and relate these new words to prior knowledge. Furthermore, engagement can serve as a measure of how well the learning activities are tailored to the child's abilities — constantly presenting tasks that are either too hard or too easy could have a detrimental effect on engagement. The results of our evaluation show that indeed the adaptive system appears to match the learning activities to each child's needs by providing a realistic yet challenging task, resulting in a reduced decline in engagement towards the end of the interaction. Gestures contribute to a higher overall engagement, which could be explained by the fact that the robot appears more active and playful in this condition, thereby stimulating the child to remain engaged.

6 CONCLUSION

The study presented in this paper aimed to explore if a humanoid robot can support children, four to six years old, in learning the vocabulary of a second language. We found that, indeed, children manage to learn new words during a single tutoring interaction, and are able to retain this knowledge over time. Specifically, we investigated whether the effects of tailoring learning tasks to the knowledge state of the learner and using co-speech gestures — both of which are strategies used by human teachers to scaffold learning — transfer to the use of a humanoid robot tutor. Our results show that the robot's use of gestures has a positive effect on long-term memorization of words in the L2, measured after one week. Furthermore, children appear more engaged throughout the tutoring session and are able to provide more correct answers when gestures are used. An adaptive tutoring strategy helps to reduce the drop in engagement that inevitably happens over the course of an interaction, by providing contingent, personalized support to each learner. By combining both methods in a tutoring session, adaptivity seems to succeed in finding the 'sweet spot' of challenging children enough to keep them motivated while gestures can add to overall engagement and support children in finding the correct answer. Therefore, gestures can form an additional tool in the toolbox of A-BKT to be deliberately employed, for example, when a reduced difficulty is deemed necessary or engagement is decreasing.

ACKNOWLEDGMENTS

This work is partially funded by the H2020 L2TOR project (grant 688014), the Tilburg center for Cognition and Communication 'TiCC' at Tilburg University (Netherlands) and the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), funded by the German Research Foundation (DFG), at Bielefeld University (Germany). The authors would like to thank all members of the L2TOR project for their valuable comments and suggestions that have contributed towards the design of the experiment. Furthermore, we are grateful to the schools, parents, and children that participated in our experiment, Elske van der Vaart for lending us her voice for the content on the laptop, as well as Sanne van Gulik, Marijn Peters Rit, and Emmy Rintjema for their help with data collection. The preliminary design of this experiment was first presented at the R4L workshop, HRI'17 [9]; we thank the attendees for their feedback.

REFERENCES

[1] Martha W. Alibali and Mitchell J. Nathan. 2007. Teachers' Gestures as a Means of Scaffolding Students' Understanding: Evidence From an Early Algebra Lesson. *Video Research in the Learning Sciences* 39, 5 (2007), 349–366. https://doi.org/10.1111/j.1467-8535.2008.00890_7.x

[2] Kirsten Bergmann and Manuela Macedonia. 2013. A virtual agent as vocabulary trainer: iconic gestures help to improve learnersâĂŹ memory performance. In *International Workshop on Intelligent Virtual Agents*. Springer, 139–148.

[3] Phyllis C. Blumenfeld, Toni M. Kempler, and Joseph S. Krajcik. 2005. *Motivation and Cognitive Engagement in Learning Environments*. Cambridge University Press, Cambridge, Chapter 28, 475–488. https://doi.org/10.1017/CBO9780511816833.029

[4] Paul Bremner and Ute Leonards. 2016. Iconic gestures for robot avatars, recognition and integration with speech. *Frontiers in Psychology* 7 (feb 2016), 183. https://doi.org/10.3389/fpsyg.2016.00183

[5] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. 2008. Gesturing makes learning last. *Cognition* 106, 2 (2008), 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010 arXiv:NIHMS150003

[6] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[7] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (2004), 241–250.

[8] Jacqueline A. de Nooijer, Tamara van Gog, Fred Paas, and Rolf A. Zwaan. 2013. Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica* 144, 1 (2013), 173–179. https://doi.org/10.1016/j.actpsy.2013.05.013

[9] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2017. Exploring the Effect of Gestures and Adaptive Tutoring on Children's Comprehension of L2 Vocabularies. In *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*.

[10] Katinka Dijkstra and Lysanne Post. 2015. Mechanisms of embodiment. 6, OCT (2015), 1525. https://doi.org/10.3389/fpsyg.2015.01525

[11] Goren Gordon and Cynthia Breazeal. 2015. Bayesian Active Learning-based Robot Tutor for Children's Word-reading Skills. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1343–1349.

[12] Lea A. Hald, Jacqueline de Nooijer, Tamara van Gog, and Harold Bekkering. 2016. Optimizing Word Learning via Links to Perceptual and Motoric Experience. *Educational Psychology Review* 28, 3 (2016), 495–522. https://doi.org/10.1007/s10648-015-9334-2

[13] Autumn B. Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin* 137, 2 (2011), 297–315. https://doi.org/10.1037/a0022128

[14] Tanja Käser, Severin Klingler, Alexander Gerhard Schwing, and Markus Gross. 2014. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *International Conference on Intelligent Tutoring Systems*. Springer, 188–198.

[15] Spencer D. Kelly, Tara McDevitt, and Megan Esch. 2009. Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes* 24, 2 (2009), 313–334. https://doi.org/10.1080/01690960802365567 arXiv:http://dx.doi.org/10.1080/01690960802365567

[16] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 67–74. https://doi.org/10.1145/2696454.2696457

[17] James Kennedy, Severin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction : Evaluations and Recommendations. *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2017), 82–90. https://doi.org/10.1145/2909824.3020229

[18] S. Leitner. 1972. *So lernt man Lernen: Der Weg zum Erfolg [Learning to learn: The road to success]*. Freiburg: Herder.

[19] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 423–430.

[20] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. *34th Annual Conference of the Cognitive Science Society* 34, 1 (jan 2012), 1882–1887. https://doi.org/ISBN978-0-9768318-8-4

[21] Manuela Macedonia, Karsten Müller, and Angela D. Friederici. 2011. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping* 32, 6 (2011), 982–998. https://doi.org/10.1002/hbm.21084

[22] Panos Markopoulos, Janet C. Read, Stuart MacFarlane, and Johanna Hoysniemi. 2008. *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter 1, 3–18.

[23] David McNeill. 1985. So you think gestures are nonverbal? *Psychological Review* 92, 3 (1985), 350–371. https://doi.org/10.1037/0033-295x.92.3.350

[24] Omar Mubin, Catherine J. Stevens, Suleman Shahid, Abdullah Al Mahmud, and Jian-Jie Dong. 2013. A Review of the Applicability of Robots in Education. *Technology for Education and Learing* 1 (2013), 209—-0015. https://doi.org/10.2316/Journal.209.2013.1.209-0015

[25] Meredith L. Rowe, Rebecca D. Silverman, and Bridget E. Mullan. 2013. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology* 38, 2 (2013), 109–117. https://doi.org/10.1016/j.cedpsych.2012.12.001

[26] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of ACM/IEEE HRI 2017*. ACM Press, 128–136. https://doi.org/10.1145/2909824.3020222

[27] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. 2016. Affect-Aware Student Models for Robot Tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 864–872.

[28] Marion Tellier. 2008. The effect of gestures on second language memorisation by young children. *Gesture* 8, 2 (2008), 219–235. https://doi.org/10.1075/gest.8.2.06tel

[29] Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in teacher-student interaction: A decade of research. (2010), 271–296 pages. https://doi.org/10.1007/s10648-010-9127-6 arXiv:arXiv:1002.2562v1

[30] Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. 2013. Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics* 5, 4 (2013), 491–501. https://doi.org/10.1007/s12369-013-0214-y

[31] Paul Vogt, Mirjam De Haas, Chiara De Jong, Peta Baxter, and Emiel Krahmer. 2017. Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in human neuroscience* 11, March (2017), 1–7. https://doi.org/10.3389/fnhum.2017.00073

[32] Lev Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.